People in Greek Documentary Papyri
First Results of a Research Project

B. Van Beek / M. Depauw (K.U.Leuven)

1 Introduction

The Trismegistos platform[1] offers an extensive set of metadata for all ancient texts from Egypt dated between 800 BC and AD 800. At present more than 110.000 texts are included, written in Egyptian (hieroglyphic, hieratic, Demotic, Coptic), Greek, and Latin, but also Aramaic, Phoenician, Arabic, and several other languages and scripts. Much of the initial work has been carried out in cooperation with the Seminar für Ägyptologie at the Universität zu Köln (Germany), in the course of the research project 'Multilingualism and Multiculturalism in Graeco Roman Egypt' (2005-2008) sponsored by a Sofja Kovalevskaja prize (2004) of the Alexander von Humboldt-Stiftung. The Trismegistos database is currently hosted by the Katholieke Universiteit Leuven (Belgium), where it is still being developed. Since October 2008, several new research projects[2] on names and identities in Graeco-Roman Egypt use the Trismegistos platform as a starting point, expanding the database's functionality to accommodate prosopo-graphic as well as onomastic information. For several languages the data collection can only be done manually in view of the lack of easily accessible full-text corpora, but for the largest group of about 50,000 Greek papyri this is fortunately not the case because the Duke Databank of Documentary Papyri (DDbDP) has been put at our disposal. This article describes the historical development of the structure of the 'People' database, the procedure implemented to automate record collection on the basis of the DDbDP, and some first results of the project.

2 The People database: history and structure

For the new projects, the collection of prosopographical and onomastic information and its integration into Trismegistos fortunately did not have to start from scratch. For prosopography we could fall back on the expertise gathered in the course of the history of the *Prosopographia Ptolemaica*, which also gave us a running start in dealing with the onomastic aspects.

---

[1] See http://www.trismegistos.org/

[2] 'Creating Identities in Graeco-Roman Egypt' (OT: K.U.Leuven), 'An interdisciplinary database of proper names in late pharaonic, Graeco-Roman and Byzantine Egypt (ca. 800BC - AD640)' (Hercules: Flemish Government). 'Egyptian names from the late pharaonic until the Roman Period. The evolution of onomastic types in a multilingual and multicultural environment' (FWO Flanders), 'Names and identities in Christian Egypt' (F+: K.U.Leuven).

2.1 The prosopographical structure (Prosopographia Ptolemaica)

As its name betrays, the *Prosopographia Ptolemaica* is a prosopography of all individuals with a title living under the Ptolemaic rule of Egypt and neighbouring territories, between 332 and 30 BC, attested in Greek, Egyptian and Latin sources, both authors and documents. The first volume appeared in 1950 in the series *Studia Hellenistica* and for nearly half a century the project was directed by the founding fathers Willy Peremans and Edmond Van 't Dack. Originally published as 'traditional' printed prosopographical lists, the digitization of this information was started in 1982.[3] The information was stored in two relational Filemaker databases, one for individuals (PER) and one for references to these individuals (REF). In some cases this typical prosopographical distinction is extremely relevant for onomastics: in the case of Ζήνων most of the over 1000 attestations (REFs) of this name in the 3rd century BC can be reduced to a single individual (PER),[4] the owner of the largest personal archive discovered from the Ptolemaic period. From 2005 onwards the digitized *Prosopographia Ptolemaica* was integrated in the *Trismegistos* platform as a subset of person-related metadata ('People').

2.2 The onomastic structure

While the prosopographical structure was already fully implemented in *Trismegistos*, the onomastic aspects only existed in an embryonic form and had to be developed at the start of our project. We will illustrate this by means of an example.

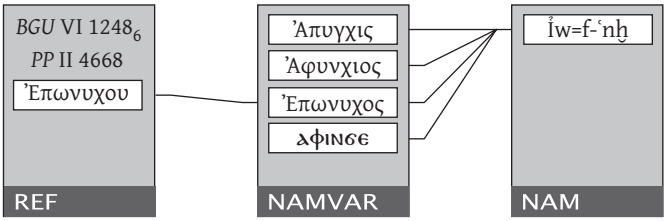
One of the guards (φυλακῖται) identified by the PP is a certain Ἐπωνυχος (*PP* II 4668), attested in BGU VI 1248. Indeed in line 6 a genitive form Ἐπωνυχου is present. The reduction of declined forms to the nominative, in this case from Ἐπωνυχου to Ἐπωνυχος, is normally made implicitly and not reflected in prosopographies nor in onomastic tools. Although necessary and essential for the automated recognition in full-text databases, we will make abstraction of it for the time being, leaving us with Ἐπωνυχος. On an onomastic level this is an attestation of a variant (NAMVAR) of a name (NAM). The distinction between variants and names may seem surprising, but is absolutely necessary in a multilingual environment. Interdisciplinary onomastic research implies a level at which the attestations of a specific name in whatever language can be brought together. This is what our NAM database aims to do, grouping variants of a specific name in all languages. After the identification of an archetypical form of the name in the etymological language (NAM), in this case Egyptian Ỉw=f-ʿnḫ, all renderings of this name in other scripts (NAMVARs) are added in a separate relational database. Thus Ỉw=f-ʿnḫ and Ἐπωνυχος are only two of the over fifty variants, including Greek Ἀπυγχις or Ἀφυνχιος as well as Coptic ⲁⲫⲓⲛϭⲉ.

---

[3] L. Mooren, 'The automatization of the *Prosopographia Ptolemaica*', in I. Andorlini, G. Bastianini, M. Manfredi, G. Menci (edd.), *Atti del XXII Congresso Internazionale di Papirologia, Firenze, 23-29 agosto 1998* (Firenze, 2001), pp. 995-1008.

[4] *PP* I 80 + add.

This results in the following visual representation:



3 Automated record collection from a full-text corpus

On the basis of this database structure we could now tackle the problem of finding a convenient and reliable way to filter out prosopographical and onomastic information from a full-text corpus of some 50,000 Greek papyri. A choice had to be made between customizing already existing named entity recognition software or creating a new system adapted to the specifics of a declined language such as ancient Greek.

3.1 The full-text corpus

The *Duke Databank of Documentary Papyri* (DDbDP),[5] an electronic full-text corpus of all published non-literary Greek and Latin papyri, ostraca and wooden tablets, was started in 1982. The texts were originally stored in beta-code, but have recently (2005) in the context of *Integrating Digital Papyrology* been migrated to the internationally recognized EpiDoc standard of TEI XML mark-up, using Unicode for polytonic Greek. This XML version was released under open access provisions in October 2008 (all content under CC BY and software under GNU GPL)[6] and may be used by other scholars for research purposes. They thus provide an almost perfect basis for computer-aided tracing of personal names, especially since the creators of DDbDP

[5] See http://idp.atlantides.org/trac/idp/wiki/DDBDP.

[6] TEI (Text Encoding Initiative) is a consortium that develops and maintains a standard for the representation of texts in digital form. XML (Extensible Markup Language) is set of rules for encoding documents electronically and is widely used in standards-compliant applications due to its simplicity, usability over the internet and strong support via Unicode for most languages. GNU GPL (General Public License) is a free, copyleft license for software and other kinds of works. Creative Commons (CC) is a nonprofit corporation dedicated to making it easier for people to share and build upon the work of others, consistent with the rules of copyright. They provide free licenses and other legal tools to mark creative work with the freedom the creator wants it to carry, so others can share, remix, use commercially, or any combination thereof. The license type BY, known as 'Attribution', allows users to copy, distribute and transmit a work, as well as to adapt its content, provided that the original provider is credited for the original creation (see e.g. http://creativecommons.org/).

have chosen to capitalize only proper names such as personal names, place names, names of deities and months in the source code. The first word of a text or a sentence has not been capitalized.

3.2 Named entity recognition: commercial vs. custom-made software

Our first option was to fall back on commercially available software applications. These can trace personal names in any given text, using named entity recognition processes (NER). Most of these NER applications, however, could not cope with ancient texts: no text recognition procedures were available for classical Greek, Latin or Egyptian texts. At best the most flexible among these existing applications could be customized to accommodate our source material, but obviously at a significant cost.

The alternative was to develop a completely new custom-made application of our own. This was feasible because the source material on which the NER would be used, in casu the DDbDP, was a standardized and in comparison with other applications rather limited corpus of xml-encoded texts, in which a very important first step for the distinction of personal names had been taken by the capitalization of proper names. This would greatly facilitate our task.

3.3 The procedure for onomastic and prosopographical filtering

3.3.1 Constructing a corpus of personal names

Our first step was thus to ask the computer to identify and filter out all words starting with a capital. This resulted in a list of some 632,000 capitalized words which corresponded to 94,840 unique forms. Not all of these were personal names: place names, divine names, and names of months had to be eliminated, as well as royal names and epithets, which form a separate research topic.

The data from the *Prosopographia Ptolemaica* provided us with a corpus of thousands of personal names which could be matched with this list of capitalized entries. An already mentioned complication was the reduction of declined forms to the nominative. This necessitated the creation of a separate database, called NAMVARCASES, in which we stored declined forms of variants of names. Because of attested anomalies and variation in the declensions, a rather full set of forms was created semi-automatically on the basis of the ending. Thus names with a nominative ending in -ης were allotted 15 declined variants (nom./voc. -ης; acc. -η, -εα, -ητα, -ην; gen. -ους, -εους, -ητος, -ηους, -ειους, -ου; dat. -ει, -ητι, -ηι, -ῃ), even if some of these were very uncommon or even impossible for some names. Since these would not match any of the attested forms in the list harvested from the DDbDP corpus, this would not be problematic.

The remaining capitalized items, for which no match could be found in this set of declined names based on the *Prosopographia Ptolemaica*, were reviewed manually. Forms likely to be nominatives were identified first, after which again all possible declined forms were created, which were then matched with the capitalized items from the xml-files. In a second stage other forms in the list identified as names were

put in the nominative and the same procedure was repeated. In the end all capitalized items were thus identified as declined forms of personal names, as something else (a toponym, the name of a ruler, a month-name ...), or as an ambiguous form.

This first phase of creating a database to recognize personal names has now been completed. At this moment (December 2009) we have 272,237 namvarcases, representing all possible declined forms of about 37,927 name variants (NAMVAR) in Greek. These Greek variants were afterwards taken together with those from other languages to create a database of currently 30,825 different names (NAM), which form the basis for the onomastic research we are doing for the project.

3.3.2 Manual review of the matching results

In the second phase the automated recognition of relevant metadata in the electronic text corpus was reviewed manually, text per text, to exclude erroneous entries or to use the context to solve ambiguities, distinguishing nominatives from genitives or even toponyms from anthroponyms.

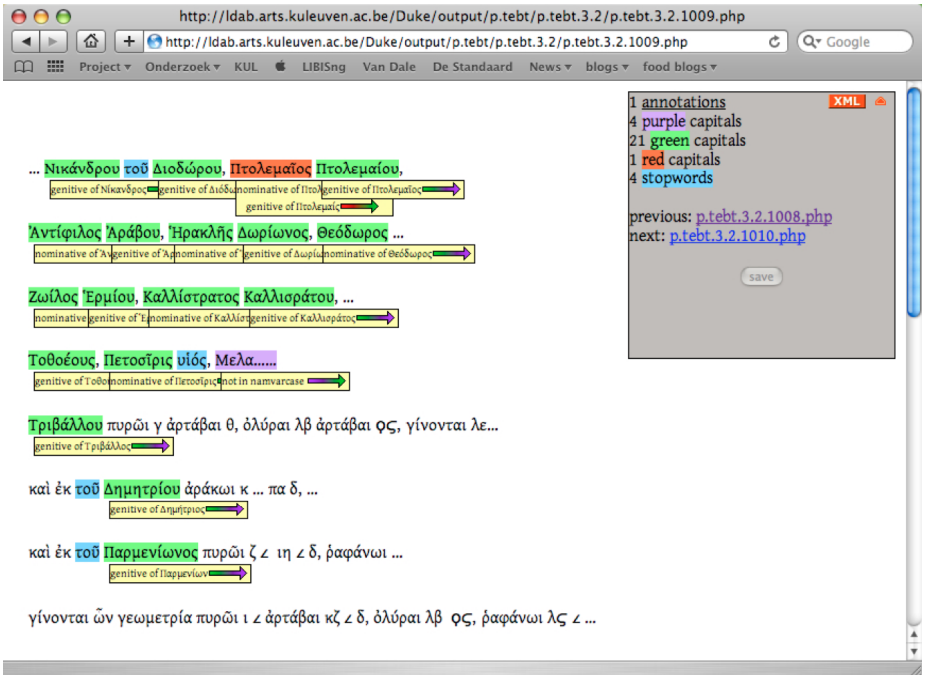


Fig. 1. Screenshot: a papyrus text stripped of (most of) its xml-markup and with capitalized words matched to the onomastics database.

The screenshot (figure 1) shows a text in which the computer has found 26 capitalized forms, 21 of which have only one match with a declined form in the database NAMVARCASE (colour-coded in green); a single capitalized form matches more than one declined form, and the computer cannot automatically identify the case involved (colour-code: red). An additional 4 capitalized words were found that were

not recognized as personal names (colour-coded in purple). In this case Μελα..... needs to be identified as a fragmentary personal name (purple to green) and the ambiguous form Πτολεμαῖος is clearly a nominative (red to green).

These choices, corrections, and additions from a human reviewer, modifying the computer-generated results, are stored in a MySQL database and added to the source text as a CSS-overlay through PHP, in order not to change the DDbDP source code.

3.3.3 Distilling genealogic and prosopographical information

In the third and last phase of the automated data entry, the personal names recognized in the DDbDP texts will be used to automatically create references to individuals. The genealogical information provided by the text will be included, and prosopographical identifications within a single text will be made as well. What distinguishes this step from the previous onomastic phase, is the added complexity of context-based information to be taken into account. Whereas declined forms of variants of names constitute an after all fairly limited set which can be used for comparison with a full-text corpus of our size, there is much more variation in person-related information. It relies on the combination of heterogeneous elements in various cases and also uses specific words. Distilling this information requires a higher degree of 'real understanding', not just character recognition and comparison. As such, it is a much more difficult task to automatize and involves more human effort, both in terms of preparing the program code for the computer and interpretating the results afterwards.

Our first step was to draw up a list of 'linking' expressions which connect or specify the various names used in a personal identification. They are words such as μητρὸς '(his or her) mother (being)', ἡ καὶ 'also known as', or πρεσβύτερος 'the older' as well as their combinations, e.g. νεωτέραν τὴν καὶ 'the younger also known as' or ἀπάτωρα μητρός 'without father, (his or her) mother (being)'. Because of these combinations and the declined forms, the 'linking' expressions are almost 1,000 in number, each with a unique numeric id. Some examples can be seen in the screenshot (figure 1), marked in blue.

In a next stage the computer identifies the strings or clusters formed by consecutive matched capitalized entries ('green') and 'linking' words ('blue'). These strings are then compared with a set of some 150 'rules' to determine which genealogical information they contain. Cases are obviously important here: a name in the nominative case followed by one in the genitive is a person with his father, while a name in the nominative followed by a name in the dative are two separate individuals. The type of 'linking' word is equally relevant: a name followed by an expression such as ἐπικαλουμένῳ 'also called' and another name is a person with his double name. An additional complication is introduced by the Latin names which increasingly appear from the 2nd and 3rd centuries AD onwards. A Greek or Egyptian name in the genitive case followed by another name in the genitive can normally be identified as the (somewhat grammatically irregular) identification of a person with his father, e.g. Λολουτος Ψενχωνσιος. But a Latin name in the genitive followed by another name in the genitive should be analyzed differently, as a

other name in the genitive should be analyzed differently, as a combined entity identifying a single individual, e.g. Αὐρηλίου Σαραπίωνος.

In view of these complications not all possible combinations can be anticipated by the rules, and again the automated decisions about the genealogical information in the strings have to be checked and if necessary corrected by human intervention. The latter is certainly necessary for an even further step in the prosopographical analysis which is combined with this manual check, in casu the identification of people mentioned more than once in a single text. Certainly for common names in longer texts only thorough scrutiny of the context can determine whether they identify the same individual or not.

At this stage the identification of individuals mentioned in different texts is consciously excluded, since it often involves looking for 'links' between documents that are somehow related. Even for a scholar with thorough knowledge of the sources a reliable identification is often tricky. The other Trismegistos databases such as Texts and Archives may well help to do this, as they make it easier to select texts with a similar background (e.g. found or written in the same place and within a certain period of time), but this task will still be very time-consuming. Intertextual identifications will therefore only be implemented systematically for the Ptolemaic period, as an update of, as well as an extension to, the *Prosopographia Ptolemaica*. For all later texts, we can only hope to identify the most frequently encountered individuals such as high officials or archive owners.

4 First results

At the time of writing [December 2009] we had just entered the second stage of the third phase: manually checking the automated genealogical interpretation of the strings and identifying individuals within a single papyrus. Some results obtained in the first two phases, however, can be presented here.

4.1 The number of name occurrences and its evolution

Comparison of the capitalized full-text corpus of the DDbDP with our onomastic databases resulted in a database with 678,849 records. Of these, 466,569 identifications were made with personal names, 364,597 unambiguous ('green'), and 101,972 ambiguous in respect to case interpretation ('red'). Another 55,878 were recognized as place names ('yellow') and 156,402 as other capitalized forms ('purple'). During manual control a team consisting of Y. Broux, S, Coussement, H. Verreth, B. Van Beek, W. Clarysse and M. Depauw checked these automated interpretations and added new names where necessary.[7] The results for personal names are shown in table 1.

[7] Up to December 2009 some 1499 non-capitalized personal names (mostly acephalous fragmentary names) were added manually. Of the 101,972 ambiguous forms ('red') 12,984 remain for the time being (not all of the texts are completely ready). The others were identi-

Starting number	364,597
added manually	+ 1,499
converted from purple to green	+ 24,209
converted from yellow to green	+ 951
converted from red to green	+ 40,415
converted from green to purple	- 34,583
converted from green to yellow	- 19,223
Final number	377,865

Table 1. Number of name occurrences recognized via automated name recognition and after manual revision.

The 377,865 personal names counted after phase 2 should of course not be considered an exact number of name occurrences, since the Latin material has not yet been included, and phase 3 will no doubt still detect errors. Nevertheless the figures are in all likelihood a close enough approximation to be used in statistical evaluations of evolutions. For onomastic and prosopographical studies, the number of name occurrences per century can replace the number of texts, which is currently the standard as indicator of the volume of available source material.[8] For this reason we have set out the number of HGV documents assigned to a specific century in Trismegistos Texts (A) against the number of name occurrences in these documents (B).[9]

Century	Documents (A)	Name occ. (B)	Av. names per text (B/A)
BC4	4	39	9.8
BC3	3,827	24,474	6.4
BC2	2,879	25,058	8.7
BC1	1,391	9,335	6.7
AD1	4,129	39,220	9.5
AD2	12,467	110,494	8.9
AD3	6,628	47,229	7.1
AD4	3,766	31,145	8.3

fied as specific cases of personal names ('green'; 41,026), place names ('yellow'; 2,130) or other capitalized forms ('purple'; 1,330).

[8] E.g. in W. Clarysse – M.C.D. Paganini, 'Theophoric Personal Names in Graeco-Roman Egypt. The Case of Sarapis', in *AfP* 55 (2009), pp. 68-89, esp. p. 72.

[9] Thus covering about 78% of the 52,875 HGV records (not including the 1381 O. Abu Mina which are not yet in the DDbDP and would thus have distorted the picture for the 7th century AD). The differences between our column A and Clarysse – Paganini, p. 72 'number of texts (HGV via Trismegistos)' is largely (but apparently not exclusively) due to the fact that they have included the entries dated to two centuries (assigning half to each of the centuries in the range).

AD5	1,046	7,170	6.9
AD6	2,849	21,585	7.6
AD7	1,578	8,721	5.5
AD8	726	5,994	8.3
TOTAL	41,291	330,464	8.0

Table 2. Number of documents assigned to a single century (A), with the number of name occurrences per century (B); the average number of name occurrences per text for each century is given in the last column.

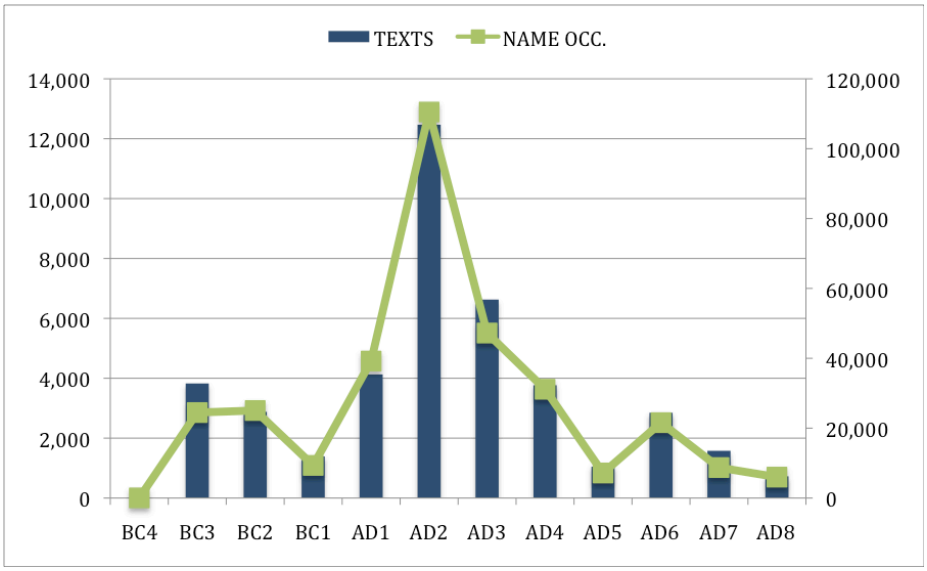


Fig. 2. Graph representing the number of texts in the DDbDP corpus (columns, scaled using the left vertical axis), overlaid with a line graph representing the number of occurrences of personal names in the same corpus (set against the vertical axis to the right).

As table 2 and figure 2 show, the evolution of number of texts per century is remarkably similar to that of the number of name occurrences per century. The working hypothesis of Clarysse –Paganini that 'the number of persons is to a certain extent related to the number of texts in each century' is thus confirmed, at least if 'number of persons' is replaced by 'name occurrences'.[10] The difference between these two term lies in the absence of prosopographical identifications in our figures and the use of multiple names referring to a single person. Examples of the latter are double names, or, statistically probably more significant, Roman style names such as Μᾶρκος Αὐρήλιος Σαραπάμμων.

[10] Clarysse – Paganini, *loc. cit.*

4.2 The number of names and the communality of rare names

An interesting recent article by G. Ruffini[11] has discussed the name frequency distribution in Byzantine Egypt, confirming the onomastic 'law' that rare or unique names constitute the largest group in onomastic data sets. In his samples (reproduced in table 3), unique names comprised between 54 and 78% of the material, while rare names (attested 5 times or less) accounted for 85 to 98% of all onomastic entries. Common names (with a frequency of 10 or more) constituted only a limited group, between 9 and 1%.

Data Set	= 1	≤ 5	≥ 10	n (size)
1. Skar Codex	54 %	85 %	9.2 %	174
2. Aphrodito Register	57 %	86 %	6.2 %	194
3. Temseu Skordon	60 %	85 %	5.3 %	207
4. Aphrodito Prosop.	67 %	90 %	6.4 %	605
5. P. Oxy. 16	70 %	89 %	5.4 %	557
6. BGU 12	71 %	95 %	3.7 %	161
7. P. Col. 8	78 %	98 %	1 %	224

Table 3. Data set reproduced from G. Ruffini, 'The Commonality of Rare Names in Byzantine Egypt', in *ZPE* 158 (2006), p. 219, table 2a.

Ruffini's data sets are not homogeneous, as he himself pointed out. They range from a single text listing tax payers in a single village in a single year (1) to the index of a volume of texts from several provenances of Egypt and ranging from the 1st to the 6th century AD (7). On the basis of the growing figures from data set 1 tot 7 Ruffini assumed that 'increasing the geographic and chronological range of the data increases the proportion of rare names'.

This can now be tested using our corpus of all DDbDP texts, as well as the conclusion that 'a data set's proportion of rare names itself indicates the extent of that data set's geographic and chronological reach, outward in space and time'.[12] In this theory, our data set should have a very high proportion of rare names. Our results using the NAMVAR database are presented in figure 3 and table 4.

[11] G. Ruffini, 'The Commonality of Rare Names in Byzantine Egypt', in *ZPE* 158 (2006), pp. 213-225.

[12] For both quotes, see Ruffini, op. cit., p. 219.

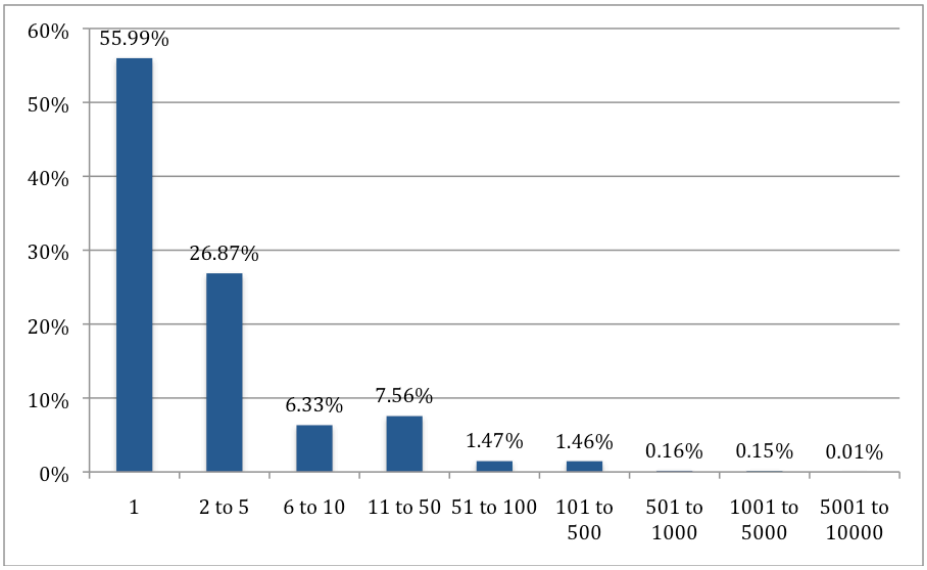


Fig. 3. Graph showing the percentage of all NAMVARs in the DDbDP corpus by frequency.

Name frequency	Number of NAMVARs	Percentage (n = 29,962)
1	16,775	55.99 %
2 to 5	8,052	26.87 %
6 to 10	1,897	6.33 %
11 to 50	2,266	7.56 %
51 to 100	440	1.47 %
101 to 500	437	1.46 %
501 to 1,000	49	0.16 %
1,001 to 5,000	44	0.15 %
5,001 to 10,000	2	0.01 %

Table 4. Number of NAMVARs with their frequency in the DDbDP corpus (absolute and percentages; total size of the sample is 29962 NAMVARs.

Contrary to what Ruffini predicted, our data set resembles that of the Skar Codex (his data set 1) more than any other, with a relatively low prevalence of unique and rare names (56 and 83%) and a high proportion (10.81%) of common names[13]. Expanding the chronological and geographical range only led to a higher proportion

[13] With Ruffini, we have taken 'common names' to be those with a frequency of more than 10; in our table of NAMVAR frequency, this is the sum of the results for the categories '11 to 50' up to '5,001 to 10,000', and amounts to 10.81%. One can of course wonder whether a name attested 11 times in over 300,000 name occurrences is really 'common'. Perhaps the border should be put elsewhere, based on the sample size. This problem was not addressed by Ruffini either (whose samples range between 161 and 605).

of unique or rare names in Ruffini's data sets, because the sets were just too small to cover the large ranges. The samples were in statistical terms not representative. This is actually just common sense: if you would try to get an idea of the name distribution by taking a small sample covering the entire world and the last millennium, you would evidently end up with a lot of unique and uncommon names. Or, to put it more positively, we can postulate another hypothesis as conclusion: the name distribution of a representative set of name occurrences or of people should approximate the values obtained by our corpus or Ruffini's Skar Codex; if the values for unique or rare names are too high and those for common names too low, the sample is not representative for the name distribution of the area and period covered. Whether this hypothesis can be falsified, should of course be further examined.

5 Future prospects

The Leuven computer-aided parsing of the DDbDP full text will result in the creation and storage of onomastic and prosopographical metadata in a relational Filemaker database. This information should eventually be made available online to the scholarly community at large as part of the Trismegistos website, in a PHP/MySQL format. Given the complementarity with the electronic corpus of the DDbDP, further integration of the various digital tools should also be contemplated. Using xml-tagging, a semantic markup within the encoding of the Greek text could be used to add related information to the actual content, a feature which can help e.g. with more effective and targeted searching of the texts.

A problem here will be how to deal with the dynamic character of both databases. Those responsible for the DDbDP should be able to continue working on the source texts in a fully independent way, without making the links with the onomastic or prosopographical information obsolete: a different reading of the verb in line 15 should not have repercussions for the link between Duke and Leuven of a personal name in line 19. On the other hand it would be optimal if Leuven were warned when onomastic corrections are implemented in the full text offered by Duke, or vice-versa. The possibilities of 'stand-of markup' assuring this exchange of information should therefore be explored: although this is a technological challenge, its development would stimulate further integration of the various research tools and thus be of real benefit to the scholarly community.